DOCUMENT RESUME

ED 206 654 TM 810 546

AUTHOR Carcelli, Larry; And Others
TITLE The Effect of Item Format on Phonics Subtest Scores of Standardized Reading Achievement Tests.
PUB DATE [80]
NOTE 22p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Grade 2; Mexican Americans; *Performance Factors; *Phonics; Primary Education; *Reading Achievement; *Scores; *Standardized Tests; *Test Format; Test Validity
IDENTIFIERS California Achievement Tests; Stanford Achievement Tests; Woodcock Reading Mastery Test

ABSTRACT

The effect on students' scores of different item formats used in standardized reading achievement tests was investigated. Thirty-one bilingual Mexican-American second grade students were given a phonics test using four different formats derived from three standardized reading achievement tests. The formats sampled identical content. A two-way analysis of variance was performed and a Newman Keuls' Multiple Range Test for Differences Between Means was calculated, since main effect for format was statistically significant. Type of format accounted for almost three quarters a standard deviation in students' test scores. Results indicated that what a student appears to know is substantially influenced by the format of the particular test used in measuring achievement. On this basis, it is recommended that greater attention should be given to the effect of test item format in selecting and administering achievement tests. (Author/AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206654

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Effect of Item Format on Phonics Subtest Scores of Standardized Reading Achievement Tests

Larry Carcelli

Utah State University

Cie Taylor

Powell Associates, Austin, Texas

Karl White

Utah State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Carcelli

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 546

Running head: Item Format and Standardized Reading Achievement Tests

Abstract

The effect on students' scores of different item formats used in standardized reading achievement tests was investigated. Bilingual Mexican-American second grade students were given a phonics test using formats derived from three standardized reading achievement tests. The formats sampled identical content. Type of format accounted for almost three quarters a standard deviation in students' test scores (statistically significant at $p < .01$, $F = 31.44$). These results indicate that what a student appears to know is substantially influenced by the format of the particular test used in measuring achievement. Greater attention should be given to the effect of test item format in selecting and administering achievement tests.

The Effect of Item Format on Phonics Subtest Scores of Standardized Reading Achievement Tests

Cronbach (1971) has defined the universe of behavior sampled by a test as including all abilities required to perceive the stimulus items and formulate responses. For an achievement test to be a valid indicator of what a student knows, it must be assured that the student gets answers correct or incorrect because of what he or she knows. If, instead of the student's knowledge of the content being tested, other variables are responsible for differentiation among student scores, correct interpretation of the results becomes more difficult and the validity of the test is called into question.

Previous Research

Previous research has investigated a number of factors which may affect how well a student does on standardized achievement tests besides what the student knows. The research summarized briefly below considered variables inherent in the tests themselves and the way questions were asked as opposed to student characteristics (e.g., locus of control, motivation, anxiety) or setting variables (e.g., time of day, student/examiner ethnicity match).

Order of Test Items

Whether test items are presented from easiest to most difficult, most difficult to easiest, or in scrambled order may affect the scores students receive. Barciskowski and Olsen (1975) reported that students feel tests are harder when the test items are presented in a decreasing

order of difficulty. Towle and Merrill (1975) showed that the order of presentation can affect the anxiety level of the test and thus influence performance. Some researchers have reported that increasing item difficulty elicits better performance (Holliday & Partridge, 1979), while others have found no effect at all (Gerow, 1980; Dambrot, 1980).

Test Item Format

The format of the test item may also influence student scores on achievement tests. Millman and Setijadi (1966) showed that Indonesian students perform better than American students on open-ended math problems while American students perform better on multiple-choice problems. Evidently, the format in which questions were asked was a determinant of students' scores.

In some instances the format of a test item decreases test scores by demanding additional intellectual skills to answer the question. Poage and Poage (1977) found that math questions containing pictures were missed more often than comparable questions without pictures on the Michigan State Assessment Test. They hypothesized that items with pictures resulted in lower scores because they required cognitive capacities involving the use of perspective that were not fully developed in younger children. In a related study Kierscht and Vietze (1975) investigated whether younger children perform better on questions that use three dimensional objects as the test stimulus rather than two dimensional pictures. Kierscht and Vietze (1975) found that preschool children performed better on the Slossen Intelligence Test which uses

objects, than on the Peabody Picture Vocabulary Test which uses pictures.

Research has also demonstrated that students perform better on multiple choice than on short answer type question format covering the same content (Kursan, Rabinsky & Pandley, 1979; Loftus, 1971). Estes and DePolito (1967) suggested that such differences occurred because the short answer format required the student to recall the information from memory, whereas the multiple choice format only required the student to recognize the correct answer. Kintsch (1970) similarly suggested that recall items required the student to both search for and retrieve information, whereas recognition items only required the student to discriminate between the presented information. Halpin and Halpin (1979) demonstrated that students performed best when the type of test content was matched with the type of item format (i.e., when recognition item formats sampled concept level questions; and recall item formats sampled knowledge level questions).

Frisbie (1973) hypothesized that "multiple choice questions limit the universe of comparisons" because they only require an individual to recognize the correct response. When answering a true-false question, however, he suggested that an individual must eliminate the incorrect response by forming a counter example from a wide universe of possibilities. Benson and Crocker (1979) administered multiple-choice, true-false and matching type questions of identical content to students

and also found a statistically significant difference in the scores for each item format. Students scored highest on the matching questions and worst on true-false questions.

Question and Answer Form

All test items require that a question be asked and an answer be given. The way in which questions are asked or answers given vary from test to test. For example, Table 1 displays four different question and answer formats from the word analysis subtests of three second grade standardized achievement tests. Differences in the form for either questions or answers may affect the scores students receive on standardized tests.

Insert Table 1 about here

Johnson, Pittleman, Ackwenker and Perry (1978) administered the vocabulary subtest of the Metropolitan Achievement test with three different types of question forms. The question forms were synonym, synonym in context, and cloze. The answer form in all cases was multiple choice. The students performed best on the synonym format in which they were presented a stimulus word and asked to find the response alternative which was closest in meaning to the word. The synonym in context format was of intermediate difficulty. In this form the stimulus word was imbedded in a sentence. Students scored lowest on the cloze format, which is a "fill in the blank format". In a nonempirical study, Roid and

Hadaynna (1978) concluded that multiple-choice item stems sampled content best when the item stems (as found in synonym in context formats) had fixed syntactical structure and when adjectives and verbs were used rather than other parts of speech as the target words in cloze tests.

Changes in the answer form may also affect the child's score on an achievement test. Grier (1975), Weber (1977) and Catts (1978) found that additional distractors decreased what the child appeared to know. Williams, Davis, Anderson and Favor (1978) and Lai-Min and Coffman (1974) administered the original and a modified version of the Iowa Test of Basic Skills to elementary school students. Williams, et al. found that an additional "all incorrect" distractor decreased scores, primarily because the students were not familiar with it. Lai-Min and Coffman added an "I don't know" distractor and also found that scores decreased. Weiten (1979) administered multiple-choice questions with either compound or single response forms and found that students scored lower on the compound response form.

Summary

Achievement tests are usually interpreted as an indication of what children know and are frequently used to make important decisions concerning the placement, advancement, and educational programming of children. If, as previous research has indicated, types of test item formats influence what a child appears to know, the types of formats used by an achievement test should be considered in selecting and interpreting achievement test scores.

Based on previous research it appears that how the test items are organized, the way in which test items are asked, and the method by which they are answered all affect students' performance on tests. The research described herein expands the findings of previous research by comparing the effects on students' scores of three test item formats taken from four commonly used standardized reading achievement tests.

Method

Procedures

The Stanford Achievement Test (level II form A), the California Achievement Test (level 12 form C), and the Woodcock Reading Mastery Test were analyzed. Four different types of question formats for testing knowledge of phonic sounds were identified. The content (knowledge of the phonic sounds) were classified into six categories: consonant sounds, vowel sounds, consonant digraphs, vowel digraphs and diphthongs, controlled vowels, and variant vowels.

A Format Test was constructed so that there were six questions of each of the four format types. The six questions of each format type tested a phonics content bit from one of the six phonics categories. All four format groups of six questions tested identical content. For instance, the students' mastery of the consonant digraph "th" was tested using all four format types. There were a total of 24 separate questions (Table 1). Format #4 was scored in two ways resulting in five scores for each phonic sound. First, format 4A was scored correct if the student pronounced the target sound correctly. Second, format 4B was scored

correct only if the student pronounced the entire word correct (this is the scoring procedure suggested by the manual).

Data Source

The Format Test was administered to 31 second grade Mexican-American students from a single school in Laredo, Texas. The second grade teachers were asked on a pre-arranged visit to supply several "average" students. The teachers were told that the students would be administered a test to determine if the question and answer formats of the standardized test that their school district had chosen to use was familiar to the students. The students were administered the Format Test in January, 1981 in small groups of six except for the last section which was derived from the Woodcock and was administered individually. During test administration, each format section was preceded by abridged directions from the test from which it was derived.

Results

The results of the two-way repeated measures ANOVA are presented in Table 2. A substantial variation in the percent of students' scoring correctly on the format types is indicated ($F\ [4,155] = 31.44$, $p < .001$). In addition, the students' scores on phonic content varied significantly ($F\ [5,155] = 11.77$, $p < .001$). There was also a significant interaction ($F\ [20,155] = 26.84$, $p < .001$). The cell and marginal means and standard deviations are presented in Table 3.

Since the main effect for format was statistically significant, the Newman-Keuls' Multiple Range Test for Differences Between Means was

calculated. The results are presented in Table 4. Group means are ranked from largest to smallest across the top of the table and from smallest to largest down the side. Pairwise differences between the means of all possible pairs of formats are presented in the matrix. Significant differences between means at $p < .01$ are indicated with asterisks. The differences between the means of formats 1 and 5 and 3 and 2 are the only nonstatistically significant differences.

Insert Tables 2, 3, and 4 about here

Discussion

The results of this study indicate that students score very differently on phonics items depending on the format used to present the items. This means that conclusions about how well a student has mastered phonics content will depend in part upon the format of the particular standardized test which is used.

In support of previous research, the students performed best on those formats which did not require the use of skills and knowledge in addition to that required to answer the question. This is directly evidenced by the differences in scores obtained on formats 4A and 4B. Format 4A was scored correct if the target sound was pronounced correctly, whereas format 4B (the method recommended by the test publisher) was scored correct only if the entire word was pronounced

correctly. Format 4B required that the students know content in addition to that of the target content.

The students scored lowest on formats 2 and 3. These formats required an extra cognitive step in much the same manner as do recall items versus recognition items. Formats 1 and 4 require that the student simply pronounce or identify a word. Formats 2 and 3 require that the student discriminate between sounds within a word, isolate the sound and identify that sound in another word. This is a far more complex task.

Since the main focus of this study was to investigate whether different format types yielded different results when testing identical content, the statistically significant main effect for phonics content and interaction with format type are not discussed further. The statistically significant interaction may be due, in part, to a sparse sampling of the domain of phonics items. In other words, future research should use more questions from each phonics content category.

Although the results are based on a fairly small sample and the Format Test did not adequately sample the content domain, they do indicate that what a student appears to know is confounded with the format of the test that is used. One explanation for this is that students may perform better on those tests which ask questions in the same way that is familiar to the student during instruction. In other words, using an unfamiliar format may well decrease the student's score by introducing an irrelevant variable and make it appear that the student has not mastered the material as well as he or she has.

School personnel have traditionally been concerned with selecting achievement tests that sample the content being taught in the curriculum. Not as much attention has been paid to selecting a test that asks questions about the content in the same manner as is done in the classroom. These results indicate that the format in which questions are asked is an important consideration for selecting an achievement test.

References

Barciskowski, R. S., & Olsen, H. Test item arrangement and adaptation level. The Journal of Psychology, 1975, 90, 87-93.

Benson, B., & Crocker, L. The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 1979, 39, 381-387.

Carcelli, L., Taylor, C., and White, K. The effect of item format on phonics subtest scores of standardized reading achievement tests. Paper presented at the Annual Meeting of the American Educational Research Association, 1981, Los Angeles, California.

Catts, R. How many options should a multiple choice question have? (a) 2. (b) 3. (c) 4. At a glance research report. Sydney, Australia: New South Wales Department of Education, 1978. (ERIC Document Reproduction Service No. ED 173 354)

Cronbach, L. J. Test validation. In Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Dambrot, F. Test item order and academic ability, or should you shuffle the test item deck? Teaching of Psychology, 1980, 7, 94-96.

Estes, W. K., & DePolito, F. J. Independent variation of information storage and retrieval processes in paired associate learning. Journal of Experimental Psychology, 1967, 75, 18-26.

Frisbie, D. A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Gerow, J. R. Performance on achievement tests as a function of the order of item difficulty. Teaching of Psychology, 1980, 7, 93-94.

Grier, L. B. The number of alternatives for optimum test reliability. Journal of Evaluation Measurement, 1975, 12, 109-112.

Halpin, G., & Halpin, G. Retention in an actual classroom setting as a function of type and complexity of tests. (ERIC Document Reproduction Service No. ED 183 622)

Holliday, W. G., & Partridge, L. A. Differential sequencing effects of test items on children. Journal of Research in Science Teaching, 1979, 16, 407-411.

Hopkins, K. D., & Anderson, B. L. Instructional module on multiple comparison techniques in research. I. A guide for selecting the "Method of Choice". Laboratory of Educational Research, University of Colorado.

Johnson, D. D., Pittleman, S., Ackwenker, J., & Perry, J. On bias diagnosis of reading difficulties - IV (Technical Report No. 464). Madison: Wisconsin Research and Development Center for Individualized Schooling, 1978. (ERIC Document and Reproduction Service No. ED 173 355)

Kierscht, M. S., & Vietze, P. M. Test stimuli: Representational level with middle class and head start children. Psychology in the Schools, 1975, 12, 309-312.

Kintsch, W. Recognition and free recall of organized lists. Journal of Experimental Psychology, 1968, 78, 481-487.

Kursan, V. K., Rabinsky, L., & Pandley, T. N. Test mode, test instructions, and retention. Contemporary Educational Psychology, 1979, 4, 211-218.

Lai-Min, P. S., & Coffman, W. E. A study of "I don't know" response in multiple-choice tests. Iowa City, Iowa: Iowa Testing Programs, University of Iowa, 1974. (ERIC Document Reproduction Service No. ED 141 371)

Loftus, G. R. Comparison of recognition and recall in a continuous memory task. Journal of Experimental Psychology, 1971, 91, 220-226.

Millman, J., & Setijadi. A comparison of American and Indonesian students on three types of test items. Journal of Educational Research, 1966, 59, 273-275.

Poage, M., & Poage, E. G. Is one picture worth a thousand words? Arithmetic Teacher, 1977, 24, 408-414.

Roid, G., & Haladyna, T. A review of item writing methods for criterion-referenced tests in the cognitive domain. Oklahoma City, Oklahoma: Paper presented at the Annual Meeting of the Military Testing Association, 1978. (ERIC Document Reproduction Service No. ED 178 562)

Towle, N. J., & Merrill, P. F. Effects of anxiety type and item-difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 1975, 112, 241-249.

Weber, M. B. The effect of choice format on internal consistency. Emory University. (ERIC Document Reproduction Service No. ED 161 940)

Weiten, W. Relative effectiveness of single and double multiple-choice questions in educational measurement. New York, N.Y.: Paper presented at the Annual Meeting of the American Psychological Association, 1979. (ERIC Document Reproduction Service No. ED 185 097)

Williams, R. L., Davis, W., Anderson, P., & Favor, K. Test format as a form of bias for black students. Journal of Non-white Concerns in Personnel and Guidance, 1978, 6, 141-147.

Table 1

Item Matrix of the Six Phonic Sounds

Tested with Four Format Types

1. Stanford Achievement Test Level II
   "Mark under the word _____."

2. California Achievement Test Level 12
   "Find the word with the same beginning sound as the word _____."

3. Stanford Achievement Test Level II
   "Mark the space under the word that has the same sound as the underlined sound."

4. Woodcock Reading Mastery Test
   "How does this word sound?"

CONSONANT SOUND			SHORT VOWEL SOUND			CONSONANT DIGRAPH			VOWEL DIGRAPH AND DIPHTHONG			CONTROLLED VOWELS			VARIANT VOWELS		
sn			u			th			ou			ar			au		
sow	snow	show	rug	rig	rag	torn	horn	thorn	sport	spout	spot	bran	born	barn	laugh	lay	leaf
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
shake	state	snack	bad	big	nut	thick	tnw	torn	loud	putt	pot	for.	from	mark	told	caught	tall
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
snip			luck			thin			about			car			caught		
ship	snack	sip	rut	lick	mop	tin	then	throw	four	cow	above	corn	arm	ran	laugh	bought	cow
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
snud			cug			sith			fout			pard			sauf		

Table 2
Standardized Question
Format Taxonomy

Form	Mode: Auditive	Mode: Visual
Language	#1 #2	
Alphabetic Symbols		#3 #4

Table 3
Standardized Answer
Format Taxonomy

Form	Mode: Writing	Mode: Verbal
Open	#4	
Multiple Choice		#1 #2 #3

Table 2

Source of Variance for

Repeated Measures ANOVA

Source	Degrees of Freedom	Sum of Squares	F	Significance Level
Phonics	5	5.16	11.77	.001
Format	4	16.48	31.44	.001
Subjects	30	82.41		
P X F	20	46.37	26.84	.001
P X R	150	13.14		
F X R	120	15.72		
P X F X R	600	51.83		
TOTAL		231.11		

Table 4

The Newman-Keuls' Multiple Range

Test of Differences Between Means

for Format Types

		Format 4	Format 1	Format 5	Format 3	Format 2
		.742	.586	.581	.409	.376
Format 2	.376	.366**	.210**	.205**	.033	
Format 3	.409	.333**	.177**	.172**		
Format 5	.581	.161**	.005			
Format 1	.586	.156**				
Format 4	.742	0				

* = $p < .05$

** = $p < .01$